SYSTEM AND METHOD FOR CLUSTERING A SET OF RECORDS

PRIORITY CLAIM

[0001] The present application claims the priority of European patent application, Serial No. 02025847.1, titled "A Method of Clustering a Set of Records," which was filed on November 19, 2002, and which is incorporated herein by reference.

FIELD OF THE INVENTION

[0002] The present invention relates to the field of data processing, and more particularly to the field of data clustering. Specifically, the present invention relates to a computationally inexpensive method of accurately clustering data records containing structured raw data that requires only two passes over the data

BACKGROUND OF THE INVENTION

[0003] Clustering of data is a data processing task in which clusters are identified in a structured set of raw data. Typically, the raw data comprises a large set of records, each record having the same or a similar format. Each field in a record can take any of a number of logical, categorical, or numerical values. Data clustering aims to group such records into clusters such that records belonging to the same cluster have a high degree of similarity.

[0004] Numerous data clustering algorithms are known. The K-means algorithm relies on the minimal sum of Euclidean distances to the center of clusters, taking into consideration the number of clusters. The Kohonen-algorithm is based on a neural net and also uses Euclidean distances. IBM's demographic algorithm relies on the sum of internal similarities minus the sum of external similarities as a clustering criterion. Those and other clustering criteria are utilized in an iterative process of finding clusters.

[0005] A common disadvantage of such conventional clustering methods is that they are computationally expensive and require a great deal of computing power. This is especially true for very large data sets.

[0006] Although this technology has proven to be useful, it would be desirable to present additional improvements. What is therefore needed is a system, a computer program product, and an associated method for an improved method of clustering that requires less computing power. The need for such a solution has heretofore remained unsatisfied.

SUMMARY OF THE INVENTION

[0007] The present invention satisfies this need, and presents a system, a computer program product, and an associated method (collectively referred to herein as "the system" or "the present system") for a computationally inexpensive method of accurately clustering of data records containing structured raw data.

[0008] Each of the data records contains a sequence of attribute values of corresponding attributes. For each of the attributes of the structured set of raw data contained in the records, a characteristic value is calculated by evaluating the attribute values of that attribute across the data records. For each of the attribute values, a deviation from the corresponding characteristic value is calculated. The attributes of each record are sorted based on the deviations to provide a sequence of attributes that is then used as a key for clustering.

[0009] The mean value or the median value of the attribute values of a certain attribute across the data records is calculated to provide the characteristic value.

[0010] The deviation of an attribute value is calculated by determining the difference between the attribute value and the corresponding characteristic value. The difference may then be normalized by dividing by that characteristic value.

[0011] The attributes of a record are sorted using the corresponding deviations for the evaluation of a sorting criterion. For example, the attributes with their corresponding deviations are sorted in ascending or descending order. The same sorting criterion may be applied for all considered records.

[0012] The clustering is performed based on the keys provided by sorting the attributes of the records. A user may select a criterion of a given number of criteria for evaluation of the keys for clustering of the data records. For example, all data records that have the same first "m" attributes are placed in the same cluster regardless of the sign of the deviations.

[0013] The clustering result is refined by searching of best matching keys in other clusters for the records of the smallest cluster. In this manner, the records contained in the smallest cluster are distributed to other clusters such that the total number of clusters is reduced. For identification of other clusters for a record in the smallest cluster, a distance measure such as a Euclids distance may be utilized.

[0014] In addition, or as an alternative to Euclids distance, gravitation may be used for reducing the number of clusters. Reference is made to the following Web site: http://www.ticam.utexas.edu/~zeyun/pick.htm.

[0015] The present system is particularly advantageous in that it provides an efficient and computationally inexpensive way to analyze the characteristics of unknown data. Furthermore, performance of the clustering method requires only two passes over the data.

BRIEF DESCRIPTION OF THE DRAWINGS

[0016] The various features of the present invention and the manner of attaining them will be described in greater detail with reference to the following description, claims, and drawings, wherein reference numerals are reused, where appropriate, to indicate a correspondence between the referenced items, and wherein:

[0017] FIG. 1 is a process flow chart illustrating a method of the record clustering system of the present invention; and

[0018] FIG. 2 is a schematic illustration of an exemplary operating environment in which a record clustering system of the present invention can be used.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

[0019] FIG. 1 shows a process flow chart for performing a method of clustering of data records containing structured raw data. Given are n records $r_1,\dots,r_n$ with k numeric attributes $a_1\dots,a_k$, where $a_i(r_j)$ is the value of the i-th attribute of the j-th record. In step 100, a characteristic value is calculated for each of the attributes. For a given attribute this characteristic value is calculated by determining a projection of the attribute values of this attribute across the records.

[0020] For example, the mean value is calculated as a characteristic value for each one of the attributes. For each attribute a_l, $l = 1,\dots,k$, calculate the mean value μ over all records as follows:

$$\mu\,(a_l) = \frac{1}{n}\sum_{i=1}^{n} a_l(r_i) \qquad (1)$$

Instead of the mean values, the median values can be calculated. The median value is calculated by determining the difference between a maximum attribute value of a considered attribute and a minimum attribute value of the considered attribute over all records divided by two. Alternatively, any other equivalent of a mean or median value may be calculated instead. By means of such mean values, median values or equivalent values characteristic values are provided for each of the attributes.

[0021] In step 102, the deviations of each attribute value of a considered record from the corresponding characteristic value are determined. For example, the deviation of an attribute value from its characteristic value can be performed by calculating the difference between the attribute value and its characteristic value. The difference may be divided by the characteristic value.

[0022] In step 104, the deviations that have been obtained for each of the records are used as a basis for sorting the attributes of this record. For example, the attributes are sorted in ascending or descending order of the deviations. In this manner, a key comprising an ordered list of attributes and associated deviations is provided for each one of the records.

[0023] Steps 102 and 104 may be carried out as follows:

- Consider record r_i.
- Consider attribute a_j.
- Calculate the deviation $\hat{a}_j(r_i)$ of $a_j(r_i)$ from the respective mean of attribute a_j using the following deviation formula:

  $$\hat{a}_j(r_i) = \frac{a_j(r_i) - \mu(a_j(r_i))}{\mu(a_j(r_i))} \qquad (2)$$

  The present system is not limited to this deviation formula; any other deviation formula may be used.
- Repeat the two preceding steps for all attributes $a_i,\ldots, a_K$ of the record r_i.
- Rank the deviations $|\hat{a}_1(r_i)|,\ldots,|\hat{a}_K(r_i)|$from the largest to the smallest, holding $\hat{a}_{l_1}(r_i),\ldots,\hat{a}_{l_k}(r_i)$. This ranking shows which attributes deviate the most from the mean of all records. For example, since $\hat{a}_{l1}(r_i)$ has the largest deviation from the respective mean value $\mu(a_{l_1})$, record r_i differs the most from all other records by attribute a_{l_1}. The largest value shows the largest deviation from the rest of the data; consequently, that attribute is very characteristic.

[0024] In step 106, the records are clustered based on the keys.

[0025] A method for performing the clustering based on the keys is to place records having identical keys into the same cluster. However, this may result in a number of clusters that is too large. Consequently, a similarity criterion is defined such that when the keys of two records fulfil the similarity criterion, the records are put into the same cluster:

Let $\hat{a}_{l_1}(r_i),\ldots,\hat{a}_{l_k}(r_i)$ be the ranking, i.e. the key, of record r_i and $\hat{a}_{l_1}(r_j),\ldots,\hat{a}_{l_k}(r_j)$ be the ranking of record r_j.

[0026] Some examples of similarity criteria are criterion A, criterion B, and criterion C.

[0027] **Criterion A:** r_i and r_j belong to the same cluster if the first m attributes of the respective keys are identical and share the same sign. For example, if the three most significant attributes (m = 3) are considered, the ranking of record r_i is as follows:

$(\hat{a}_7(r_i),\hat{a}_2(r_i),\hat{a}_3(r_i),\hat{a}_9(r_i),\ldots) = -1.17,0.95,0.87,0.56,\ldots$

and the ranking of r_j is

$(\hat{a}_7(r_j),\hat{a}_2(r_j),\hat{a}_3(r_j),\hat{a}_1(r_j),\ldots) = -1.46,1.09,0.89,0.88,\ldots.$

The records r_i and r_j belong to the same cluster, as the first three attributes of the keys are identical as well as the signs of the values.

[0028] However, if the ranking of r_k was $(\hat{a}_7(r_k),\hat{a}_2(r_k),\hat{a}_3(r_k),\ldots) = -1.46, -1.09, 0.89, 0.88, \ldots$, the r_i and r_k would belong to different sections because the signs of the second most distinguishing attribute $\hat{a}_2$ had a different sign compared to the respective value of record r_i.

[0029] **Criterion B**: r_i and r_j belong to the same cluster if the first m attributes are identical. For example, considering the previous example, records r_i and r_k would belong to the same section, even though the sign of the second most distinguishing attribute is different.

[0030] **Criterion C**: r_i and r_j belong to the same section if the same attributes appear on the first m positions with identical signs. This criterion ignores the order in which the attributes appear. For example, if m = 3, r_i as before and the ranking of r_j is $\hat{a}_2(r_j),\hat{a}_3(r_j),(\hat{a}_7(r_j),\hat{a}_1(r_j),\ldots,) = 0.72,0.68,-0.42,0.37,\ldots$ then a_2,a_3 and a_7 are identical and share the same signs. This criterion can be varied by ignoring the signs.

[0031] The resulting clustering may be further refined by reducing the number of the clusters. For example, it may be desirable to dissolve a cluster having a small size, i.e., having a small number of records. This may be accomplished by means of the following iterative process:

- Rank the clusters by size.
- Select the smallest cluster.
- For each record of the cluster, find the one of the larger clusters that matches most of the significant attributes. If more than one cluster should be considered, either choose the largest of these clusters or use some kind of distance measure to find the nearest cluster.
- Repeat until the desired number of clusters has been reached or if the similarity of records and clusters is too small.

[0032] FIG. 2 illustrates a data processing system 200 in which a system and method for clustering a set of records according to the present invention may be used. Data processing system 200 comprises a database 202 for storing records of structured data. Each of the records has attribute values $a_1,\ldots,a_k$. Each of the records has an associated data field for storing a key for that record and a data field for storing a cluster identifier. Initially the key and cluster data fields are empty.

[0033] In addition, data processing system 200 comprises a characteristic value module 204 for calculating of characteristic values for each one of the attributes. The calculation of the characteristic values may be performed as explained with respect to step 100 of FIG. 1.

[0034] Further, data processing system 200 comprises a deviation module 206 for calculation of the deviations of the attribute values. This calculation may be performed in accordance with above equation (2).

[0035] Sorting module 208 of the data processing system 200 sorts the attributes of the data records by applying a sorting criterion to the deviations of the corresponding attribute values. In this manner, a ranking of the deviations may be obtained for each record. The sorting may be performed as explained with respect to step 104 of FIG. 1.

[0036] Further, data processing system 200 comprises criteria A module 210, criteria B module 212 and criteria C module 214 for application of the respective criteria A, B and C. The criteria A, B and C are described above with respect to FIG. 1.

[0037] Further, data processing system 200 comprises a user interface 216. By means of the user interface 216, the tabular data contained in database 202 may be visualised. Furthermore, a user may select a subset of the records contained in the database 202 for performing a clustering operation. Before the data clustering is performed, the user selects one of the pre-defined clustering criteria A, B or C. Alternatively, the user may define a user specific clustering criterion.

[0038] The data clustering is initiated after the user has selected the set of records of the database 202 on which the data clustering is to be performed and after a criterion for data clustering has been selected or specified.

[0039] The characteristic module 204 is invoked to calculate the characteristic values of the attributes. The deviation module 206 is invoked to calculate the deviations of the attribute values from their corresponding characteristic values. By means of sorting module 208, the attributes are sorted to provide a key for each one of the selected records. The desired module for applying the selected criterion is invoked, i.e., criteria A module 210, criteria B module, or criteria C module 214. Alternatively, a user specified module may be invoked to apply the user specified criterion. As a result of the application of the selected or specified criterion, the selected records are clustered. Records that are placed into the same cluster are assigned the same cluster identifier; this cluster identifier is entered into the corresponding data field within database 202.

[0040] It is to be understood that the specific embodiments of the present invention that have been described are merely illustrative of certain applications of the principle of the present invention. Numerous modifications may be made to the present system, method, and service described herein without departing from the spirit and scope of the present invention.